UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report # 101

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Methodology, and the Statistician's Responsibility for BOTH Accuracy AND Relevance,

5. TYPE OF REPORT & PERIOD COVERED: Technical Rept.,

6. PERFORMING ORG. REPORT NUMBER: TR-101

7. AUTHOR(s): John W. Tukey

8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0453

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Princeton University, Princeton NJ

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research (Code 436), Arlington, VA 22217

12. REPORT DATE: Dec 1975

13. NUMBER OF PAGES: 14

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Presented at the annual meeting of the American Statistical Association in Atlanta, Georgia, August 1975. John W. Tukey is also Associate Executive Director-Research, Beel Telephone Laboratories.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

level or change, least squares, fit PLUS residuals, Phillips curve, patch maps, standardization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two contrasting themes run through this paper. The negative theme that we should do only what is very simple and obviously "exact". This has brought us to many "pretty passes". The positive theme that we must always think in terms of "data EQUALS fit PLUS residuals" and always try to extract more from the residuals.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Methodology, and the Statistician's Responsibility for BOTH Accuracy AND Relevance

John W. Tukey

Princeton University* and Bell Laboratories
Princeton, New Jersey 08540 and Murray Hill, New Jersey 07974

ABSTRACT

Let us be prepared to measure what is needed for policy guidance, even if it can only be measured poorly.

Let us recognize the importance of change, not leaving it to be a consequence -- a poor relation -- of the measurement of level. If the price is first a preliminary value and then a revision, let us pay it. Measuring charge and measuring level are not the same thing in active practice (though they may be for history).

Let us face up to the potential inadequacies of standardization for broad groups. Let us learn about further corrections. Let us, at least, publish bounds on the sizes of these further corrections. We need information on death certificates about usual places of residence, 5, 10 or 20 years before death. We can adjust observations for the most crucial effects of several/many concomitant variables by "smear and sweep", we can combine adjustment for smooth and stratum-wise dependence on a single variable. We can superstandardize.

We could usefully give more attention to: the proper abscissa for the Phillips curve, the economic analogs of physical chemistry, sampling of student histories in school systems.

So-called "statistical maps" do not deserve so honored a name. "Patch maps" is more appropriate. We can, and must do better: by assigning values to centers rather than areas, by learning to adjust for area compositions, by bringing in spatial smoothing.

Two contrasting themes run through all the topics above. The negative theme that we should do only what is very simple and obviously "exact". This has brought us to many "pretty passes". The positive theme that we must always think in terms of "data EQUALS fit PLUS residuals" and always try to extract more from the residuals.

*Prepared in part in connection with research at Princeton University sponsored by the Office of Naval Research.

This talk was given at the annual meeting of the American Statistical Association in Atlanta, Georgia, August 1975.

Methodology, and the Statistician's Responsibility for BOTH Accuracy AND Relevance

John W. Tukey

Princeton University* and Bell Laboratories
Princeton, New Jersey 08540 and Murray Hill, New Jersey 07974

The theme of this meeting is statistics in the service of policy, public or private. If we are to serve policy, we must be responsive to its needs -- often to something between its true needs and its perceived needs. This loads us down with a variety of responsibilities that we may not like, including a responsibility for understanding what policy needs really are.

I would have liked to sprinkle this talk with concrete examples where the statistical machine has chosen to measure what is easy to measure rather than what would be more helpful, even though it cannot yet be measured well. But I spend my time too far from the fray, either in Government or in Business, to have a good list. So I will have to concentrate on some general points, tied closely to methodology.

That we have failed to measure things that would be more useful because the measurement would be inaccurate is, I believe, beyond question. Even in the halls of this meeting, a friend heard a discussion summarizable as "it's possible to do it right, but it will take years, so let's hang fire till then." This may have been the correct conclusion, but it could not be known to be without two other assessments: How urgent is the policy need for this measurement? How well would a poor measurement serve policy needs? These I believe were missing in this instance, as they so often have been. We cannot afford an ivory tower attitude, an attitude that "we don't want to try to measure anything where we cannot be proud of the measurement process."

Such a spirit can produce good measurement. But when the right thing can only be measured poorly, it tends to cause the wrong thing to be measured, only because it can be measured well. And it is often much worse to have good measurement of the wrong thing -- especially when, as is so often the case, the wrong thing is IN FACT used as an indicator of the right thing -- than to have poor measurements of the right thing.

*Prepared in part in connection with research at Princeton University sponsored by the Office of Naval Research.

This talk was given at the annual meeting of the American Statistical Association in Atlanta, Georgia, August 1975.

DIFFERENCES

Level or Change

The clearest example I know of where our profession has frequently resisted using its knowledge consists of all the economic series where change is more important than level -- and what economic series is there for which this is not the case? If you tell those who do not know what the unemployment rate is today that in 1980 it will be 6.5%, they will not know whether to be sad or glad. If you tell them it will be 1% or 2% or 3% less than today, they will know how they want to react.

Our profession has given technical attention to the measurement of change. Our unemployment estimates come from samples where a household appears over adjacent months and then about a year later. We could do much more of this, and go even farther. Occasionally, we do.

But any discussion of focussing our measurement on change, something we have the technology to do, runs into the argument that concentrating on change means a weakening of our measurements of level, means a possibility of drift, means a possibility -- horrors -- of having to make a publicly visible adjustment -- a correction.

The essence of this argument is clear. We shall not measure what is most useful, because then the sampling error, which all professionals know is there, might have to be revealed to the world.

This is a sad argument -- it might even be taken to be a resort to professional pride rather than national need as a criterion. I know my colleagues too well to believe that this is the whole story -- there are real dangers in giving better values that look worse -- but that such an explanation can even be contemplated is not an occasion for pride.

Two-Period Panels

Let us look at an oversimplified case, where technology and issues are both clear. Suppose we are estimating something -- those who wish may think of unemployment -- every month, and that every sample unit is included in exactly 2 adjacent months -- there is the oversimplification.

In month i, then we have results

$$Y_{iB} \text{ and } Y_{iF}.$$

The former for exactly those sample units that gave $Y_{(i-1)F}$ and the latter for exactly those sample units that will give $Y_{(i+1)B}$. To assess the series in month i, the best we can do ought to be

$$\frac{1}{2}(Y_{iB} + Y_{iF}).$$

And what of the change from $i-1$ to i? The two bases upon which we can assess this change are

$$Y_{iB} - Y_{(i-1)F}$$

and

$$Y_{iF} - Y_{(i-1)B}$$

the first of which involves exactly the same sample units, while the second involves sample units that appear only once during the two months. Their variances will be in the ratio of $1-r$ to 1, where r measures the correlation for one sample unit from one month to another, which may easily be as large as 0.9.

Thus the best current measure of change is

$$\frac{(Y_{iB} - Y_{(i-1)F}) + (1-r)(Y_{iF} - Y_{(i-1)B})}{2-r}$$

with variance proportional to

$$\frac{1^2(1-r) + (1-r^2)1}{(2-r)^2} = \frac{1-r}{2-r}$$

as compared with the variance, proportional to

$$\frac{1-r+1}{2^2} = \frac{1}{2} - \frac{r}{4}$$

of the difference of the best individual values.

For $r = .9$, a large but not staggering value, these variances are in the ratio of 3.07 to 1. for $r = .5$, surely quite moderate, the ratio is 1.12 to 1. How large differences in quality are we prepared to accept -- if that only purpose is to maintain an appearance of consistency?

A Possible Consequence

What if we focus on change, month after month, and link the changes together? If we sum

$$\frac{(Y_{2B} - Y_{1F}) + (1-r)(Y_{2F} - Y_{1B})}{2-r}$$
$$\frac{(Y_{3B} - Y_{2F}) + (1-r)(Y_{3F} - Y_{2B})}{2-r}$$
$$\frac{(Y_{4B} - Y_{3F}) + (1-r)(Y_{4F} - Y_{3B})}{2-r}$$
$$\frac{(Y_{5B} - Y_{4F}) + (1-r)(Y_{5F} - Y_{4B})}{2-r}$$

we get

$$\frac{(Y_{5B} - Y_{1F}) + (1-r)(Y_{5F} - Y_{1B})}{2-r}$$

PLUS intermediate terms, specifically

$$\frac{r}{2-r}(Y_{4B} - Y_{4F} + Y_{3B} - Y_{3F} + Y_{2B} - Y_{2F})$$

As we add more and more changes the list of intermediate terms gets longer and longer, and hence more and more variable.

As a result the sum -- base value PLUS all monthly changes to date -- tends to drift. Sooner or later the difference of current best estimates becomes better. Summation of changes over many periods is not the way to get accurate levels.

Least Squares

Some of you will wonder how such things can be -- at least if least squares, are used -- because they will recognize that, under least squares the best estimate of any linear combination is the same linear combination of the best estimates of the components. Is this a paradox?

Not at all, for we are not in the convenient least squares paradigm:

- get all the data first,

- then analyze it.

We have to report on this month, before next and later months are in. As more data accumulates, the least squares solution for past months changes. If we were wholly free to readjust past months, there would be no conflict between measuring change and measuring level, at least for least squares (if the occurence of strings is weak enough to make least squares tolerable).

In an Ideal World

In an ideal world, what should we do? It seems to me that we ought to teach all those who respond to our economic series a simple and difficult lesson: The best estimate of change is often not the change in the current best estimates. In particular it may be that the best estimate last month was 111, this month 114 yet the best estimate of change is +2 NOT +3.

I don't know whether we can educate our consumers to the point where not only will they accept estimate of change $\neq$ change of estimate for themselves but they will even not be bothered by partisan political debate in which each debater focuses on the side of the inequation that supports his own position. But it is not clear to me that we should not try.

In an even more ideal world, of course, we would meet the situation by saying: "Last month we reported 111, we now know that 112 is more precise, this month we report 114, corresponding to a change of +2." This, in a least squares context, would be even more precise. Would our consumers accept this? If there are only few of them it seems to me that they might.

We ought to seek out places where we dare try such reporting. We ought to learn how to say such things so that they will be accepted with the least friction and lost motion.

Preliminary Series

More and more series, I suspect, are now appearing in preliminary and final forms. In almost every case where this happens, it would seem to me that the preliminary estimate should be calculated in the form: last period final PLUS best estimate of change and that all those concerned in gathering and working with data for the preliminary estimate should be brought to believe that their purpose was the estimate of change.

Preliminary estimates are usually a consequence of incomplete data. An emphasis on change would lead to more careful specification of what was the corresponding figure last month for each figure so far available this month and on what allowance -- IN CHANGE TERMS -- ought to be made for the unavailable figures. This ought to improve the preliminary figures as measures of change -- the purpose for which they are almost always wanted in a hurry.

If any of you wish to accuse me of asking us to treat current information on our economy in the spirit that current information on the vote is treated on election night in projecting election results, I will not object. Election night projection is real-time statistics at its acme -- the results come in before the projections are forgotten (except, in 1972, for the Senator from New Hampshire) and there is serious competition for timeliness and correctness. Economic statistics are slow real-time statistics by comparison, slowed by a factor of perhaps 50 to 100. Why should we not look to the attitudes and approaches that have been hammered out in the most competitive arena?

Close

Somehow we ought to move further toward measures of change -- how far we can go may have to be learned from experience -- but we ought to be pushing steadily against the practical limitations. We need to measure well what is most important for policy, even if doing this makes us trouble.

I understand that this is in fact done for the current estimates of monthly retail sales, where correlations are not just .9 but are more nearly .95 or .99. Here we have been responsive to need, as we should. I wonder, however, how thoroughly the emphasis, both within the organization and in the official releases, is on change for the preliminary and on benchmarking a level in the final.

Some of my friends who know Washington bureaucracy from the inside feel that I have been more than a little unrealistic in this discussion, saying that people get promoted in the bureaucracy for NOT making mistakes, that the difference between imprecision and inaccuracy is not recognized, so that absence of precision is taken as inaccuracy and failure, that admission of imprecision is incredibly dangerous in what is essentially an adversary process.

They go on to say we must remember how unpopular facts are when they oppose a politician's opinion, that given the various attempts of not many years ago to subvert the federal statistical process, the real problem may be our failure to visualize a strategy for making more effectively available the achievable information without imperiling either individual careers or the reputation of the federal statistical process.

STANDARDIZATION

Let me now turn to a quite different area, to a different sort of statistical technique applied, for the most part, to a different sort of data. The most used statistical technique to which most academic statisticians have given little attention is standardization.

I believe it is time for all of us to think a little more about this subject, especially about the aspects that have received least attention.

Let me outline two examples:

- standardization of death rates as in each year's Vital Statistics of the U.S.

- standardization of death rates following surgery, say by hospital, as in the National Halothane Study [1, pp. 358 ff.].

A Simple Example

In the first example, if death rates are to be reasonably comparable from year to year, we must take account of the changing age structure of the population. A simple way to begin this for white males, for instance, is to use estimated current population, in ten-year age bands, and enumerated current deaths, in the same bands, to calculate apparent death rates for each ten-year band. If we then take a standard distribution of white males over these bands -- as of the 1970 Vital Statistics publication, the U.S. 1940 population was taken as standard -- we can combine these age-band death rates to an overall standardized death rate.

Clearly, such a procedure improves comparability.

Equally clearly, it does not tell the whole story.

Death rates are not, of course, constant at one value from age 50 to age 59 -- and at another from age 60 to age 69. There are changes within 10-year bands. If the distribution of ages within a 10-year band changes, even if the death rates for each detailed age remain the same.

We have adjusted -- as we almost always do -- for broad categories, without going on to make an, at least plausible, further adjustment for the fact that the categories are not narrow. This is something we can do better if we wish. Why don't we?

It may well be that the additional corrections would be small. When this is so, however, we need to know that it is so. Either we should get further adjusted figures, with a statement that further adjustments were all small, or we should get initially-adjusted figures with a statement about how large the further adjustments would be -- at least for an honest sample of the figures given. Only thus can we be sure that, when the additional adjustments are not small, we will know it and have a chance of having adequately adjusted figures.

My concern here is not really with the national age-adjusted death rate, something that may well deserve rather more attention than it gets. My concern is directed more toward the more detailed comparisons that are now becoming available, such as the National Cancer Institute's mammouth work [2] which gives 1950-1969 age-adjusted death rates for 34 cancer sites for every county in the U.S. -- by sex and color. This is a large labor, offering us much raw material for analysis, but I cannot help wondering how much difference would be made -- at least in some cases -- by further adjustment for broad categories. If we are to disentangle the messages of geographic differences in cancer death rates, we are likely to have to face such questions.

The only two reasons for confining oneself to simple age-adjustment seem to be

- it is less work,

- it requires no VISIBLE assumptions.

In this era of the computer, the first seems hardly enough. And the second, as we may recognize, is purely a matter of visibility, not truth.

If we ask what is required for simple age adjustment to be satisfactory, the answer is "nearly enough constant shape of distribution of ages in each band separately." Yet who explains this when talking of age standardization? As long as we do not tell this truth, age adjustment seems so simple and obviously true -- let us not rock the boat.

At least among professional circles, there ought to be clear ideas of how large -- and how geographical*ly* systematic -- the effects of further age adjustment might be. (Consider Pinellas County, Florida with its high concentration of retirees -- and, presumably, still the highest median age of any country in the country. Is its distribution of ages between 60 and 65, or between 65 and 70 similar to more typical countries? Who knows off-hand?)

Death Certificates

Given the rise of deaths from possibly delayed causes -- and the precipitous fall in deaths from immediate causes, including infectious disease -- is it not time to reconsider what location information we try to get on the standard death certificate? To ask for more information is to raise new problems of incompleteness, but again perhaps we should face difficulties in the hope of better answers.

One guess as to what we might best add is information on the deceased's residence at one or more earlier periods, perhaps 5, 10 or 20 years before death. Without something of this sort, and given the mobility of our population, I find it hard to see how we can extract vital information about the effects of environmental exposure -- or even the effects of geographic differences in health care -- from our death certificates.

Yet today these are the basic issues, these, not the impact of infectious disease, are the reasons why a death registration system can effectively support public policy in the decades ahead.

Again it will be a harder job to bring measurement closer to policy. Again much wisdom is needed to consider the alternative routes by which we might begin to gather such data.

A More Complex Example

One aim of the National Halothane Study [1, pp. 287 ff.] was to compare the rates of death during 60 days after surgery for operations carried out under various kinds of anesthetic. The study was initiated by suspicions of a widely used anesthetic called halothane. Toward the end of the study, one statistician conveyed his understanding of the data to a committee mainly composed of anesthesiologists by asserting that he would not let any of his immediate family be operated on without a signed paper saying that halothane WOULD be used. While that was perhaps an overreaction, I would not like any of you to go away with a negative feeling about halothane. Today, the medical professions seem, to statisticians, to have a reasonable view of the pros and cons of halothane. Now for the statistical issues.

When one started to compare death rates after operation *ceteris paribus* there were many *ceteris* including:

- type of operation,
- age and sex of patient,
- physical status of patient,
- length of operation.

It was clearly desirable to standardize for all these things together. One difficulty was obvious, if we divided the cases into cells in accordance with even a moderately fine classification on these variates there would be more cells than deaths -- and since there were only 800,000 cases in the retrospective study, there could even be more cells than cases. Ordinary techniques of standardization could not be used, even though age of patient was in 10-year bands, much too wide for either effectiveness or efficiency -- since, at the older ages, the death rate doubled every decade.

One cure for this problem was a technique called "smear and sweep" in which standardizing variables were introduced one at a time, each being combined with those already considered before the next was introduced. I refer you to the National Halothane Study for details, mentioning the technique here only to show that one can standardize for SUITABLY SELF-SELECTED combinations of more standardizing variables than one would think.

Combined Age Dependence

If we are to work with age in blocks, there is no reason why we should look at raw death rates. If, as in the higher ages of the National Halothane Study, we have a death rate that is -- even quite crudely -- exponential, we can adjust our estimates of the number exposed in an age band to a mid-age, assuming a specific exponential dependence, and then adjust further, as needed, by factors for age bands. (Had we had ages in individual years, rather than decades, the precision of the National Halothane Study's results could have been appreciably increased.) Doing this sort of thing would increase the difficulty of explanation somewhat, but could make the results appreciably more precise -- and often correspondingly more useful.

Further Adjustment

In cases like the county-by-county cancer death rates, the possibilities of adjusting for demographic variables in addition to age, sex and color seem to me to be important. If there is an urban-rural gradient, say in the death rate, it seems to me that recognizing this fact AND removing a suitable fit, so that geographic detail is not encumbered by an easily described trend, could be very important.

The basic lesson is

data = fit PLUS residuals

promotes understanding when BOTH fit AND residuals are examined CAREFULLY -- deserves more attention in many applications.

Taking the view that all we want to do is say "what happened" in each of 3000-odd counties is not going as far as we should go. The price of going further is more complex patterns of analysis and less simple-minded description. Let us plan to pay the price.

Superstandardization

The National Halothane Study introduced another technique. For each of 34 hospitals there were available

- a raw death rate, and
- a standardized death rate,

whose ratio

$$\text{SMR} = \frac{\text{raw death rate}}{\text{standardized death rate}}$$

indicated how much adjustment to the raw death rate for that hospital was accounted for by standard, rather study (except for smear and sweep) methods of standardization. A plot of

log standardized death rate
AGAINST
log SMR

was made one afternoon (34 points, one per hospital), and a further regression, with a slope of about 0.6, manifested itself.

The question of just how to interpret such additional regressions, here

log raw death rate constant* + 1.6 log SMR

instead of

log raw death rate constant** + 1.0 log SMR

is not an easy one. Interpretations in different fields of application would seem likely to differ in character. But wherever such "superstandardization", such additional regression, sops up important parts of the variation, it seems important to make such a fit, and then -- as always -- to look hard at both the fit and the residuals.

Suppose we find that, when standardizing for age, there is additional regression on relative age, however expressed. We need to ask ourselves carefully why counties with older populations, for instance, behave older -- or younger -- than they are -- and we need to exclude this additional regression before looking at geography.

Close

What have we been asking for in connection with standardization? Mainly three things:

- facing up to corrections for adjustment by broad categories.
- flexibility in using such devices as combined age adjustment or smear and sweep.

- empiricism in using either external variables (perhaps demographic, perhaps economic, perhaps ?) or superstandardization in moving toward a

fit PLUS residuals

stance, with both being examined carefully.

Each of these comes down to recognizing complexity in the world, to making results somewhat harder to explain but, also, hopefully, to making much larger increases in the usefulness of the results.

The Phillips Curve

Economists have shown increasing interest in the "Phillips curve" the believed-in relation between unemployment and rate of inflation -- a relation that is believed to have so shifted that a given inflation rate is now consistent with a higher level of unemployment. Is this shift a change in structure -- or is it just our failure to measure the discomfort of unemployment -- rather than its rate?

Unemployment insurance is only one of the medianisms that makes unemployment less painful to individuals and households that it would have been 45 years ago. If we could measure what fraction of the people were hurt, to a given degree, by their unemployment, we might find the comparison across decades quite different than if we measure only how many are unemployed.

Might it not be that the real relation is between pain of unemployment and rate of inflation? And if this is indeed so, what are the policy consequences? Who wants higher inflation? Who wants higher pain of unemployment? A very uncomfortable dilemma that we do not know whether we are in -- perhaps only because we have failed to try to measure the more difficult quantity.

Thermodynamics or Physical Chemistry?

To one brought up a chemist, as I was, classical economics is a clear analog of thermodynamics, it tries to tell us what will happen, but not how it will come about. The fact that a rubber band must cool off when it is stretched is easy to establish by thermodynamic arguments, but to understand the mechanism by which this really happens is a matter of molecular structure, broken bands, physical chemistry. Just as real chemical processes are likely not to be perfectly reversible, so most markets do not involve perfect competition. As the Wall Street Journal [3] quoted Paul Samuelson on last Friday "Over the long run, the Japanese will come in and new competition will spring up. But by that time, many executives will be retired."

Is it not likely to be generally true that the challenges of economic measurement for the decades ahead, both for the large firm and for the nation, are the challenges of measuring characteristics of the details of the processes of re-adjustment -- and not just measuring the analogs of thermodynamic quantities. The economic analog of physical chemistry, with its detailed mechanisms, may well be the problem area of the economic statisticians' future. And think how hard it will be to measure such things, particularly at first. But then plan to go ahead and measure them.

Metropolitan School Systems

Measurement of individuals has long been characteristic of education at all levels, both to answer "which ones learned enough?" and, to a lesser extent, to answer "which ones learned best?" Over the last dozen years, measurement of education's product -- for which the "in" word is "assessment" has been becoming more and more respectable, at national, state, and even some local levels.

The next step is presumably, the measurement of the process -- something quite different from measurement of the product, especially because who knows where the product learned what he or she appears to know. (Measurement of the process at the level of the individual teacher meets strong opposition, opposition that is the more justified the less that we are able to tell what a teacher either might or ought to succeed in doing.)

Presumably, then, the first real step toward measuring the process will be the careful longitudinal study of SAMPLES of our school children. Really extensive data on all children in a large school system is probably unmanageable for the near future -- and too expensive as well. but extensive data on samples of something between 1% and 10% in a large system ought to offer us considerable insight into the process. We might someday answer questions as: What is the relation of absenteeism and rate of improvement of performance? Does it matter whether the absence is due to sickness or playing hooky? What is the return from specific sorts of curriculum enrichment?

Any of us could design an adequate sample -- quality and efficiency would of course vary with our individual experience and background. But who among us has thought long enough to have high confidence in his or her methods of measuring the things that need to be measured? Again we need to move ahead.

PATCH MAPS

I like displays, graphs and maps as much as the next man or woman -- probably much more. But I am coming to be less and less satisfied with the sort of maps that some dignify by the name "statistical map" and that I would gladly revile with the name "patch map".

At issue is the sort of map in which each county -- or each state -- or each piece of some other subdivisioning -- is colored or shaded to reflect its average characteristic. Thus the map, though perhaps not a collection of shreds and tatters, is a collection of patches, whose shape, size and location reflect the civil divisions or census subdivisions, whose data we have chosen to map.

Undoubtedly the excuse for doing just this is again that it is simple, and no one can claim you have done anything wrong. but it is a very real question whether you have done anything right, as a few extreme examples easily show.

There are exceptions, of course. Thus patch maps of taxable property per head, or per school child, do reflect a common reality for all those who live in the same patch. and I am sure you can find a few others. But how many?

An Example:

My favorite example is Washoe county, Nevada. And since the 1962 City and County Data Book was at hand, we will see data from the 1960 Census for population and area:

Division	1960 Population	1960 Area	1960 Density
Washoe County	84,743	6281	13.5
Reno	51,470	11.8	4362
Sparks	16,618	2.7	6155
Rest of County	16,655	6266.5	2.66
(Adjacent*)	(5708)	(9702)	(.59)
(Counties)	(3199)	5993	(.53)

Clearly, assigning a single population density to Washoe County is near nonsense -- and the same would be true, to a somewhat lesser degree, for assigning a single value to anything that varies noticeably from urban to rural. the average population density, of 13.5 per square mile, is unrepresentative in one direction for individuals, more than 80% of whom live in places with more than 4000 per square mile -- and is unrepresentative in the other direction for area, more than 97% of which averages less than 3 per square mile (and probably 90% of which has only a few tenths of a person per square mile. Adjacent Humboldt county, less the county seat of Winnemucca, has less than 1/4 of a person per square mile.)

In thinking about this example, remember that Washoe county has an 192-mile N-S border with California, and extends 170 miles North of Reno and Sparks. Only about one-third of the county's area is within 50 miles of these population centers.

This is an extreme example, but something of the same sort happens in almost every county. Mercer County, New Jersey, where I live, was 40% Trenton in 1960. yet much of the county was very far from being like Trenton. Bristol County, Massachusetts, where I was born, was more than 60% New Bedford PLUS Fall River PLUS Trenton in 1960, yet Westport, where I summer, operates effectively with Selectmen and Town Meetings. Maps of most quantities of direct interest which assign average values to the wholes of counties, thereby lie, lie, lie.

What To Do?

It is not that we are barred from taking action about such lies, for there are other things we can do. Some of them are:

- assigning county values to county centers (of population, area or what have you) rather than to county patches.
- smoothing such county-center values and contouring the resulting smooth.
- studying the dependence of the quantity that concerns us upon, perhaps such variables as population density, population potential, and per capita income, and then examining both fit and residuals carefully. (Smoothing and contouring the residuals may really tell us something.)

In many places, and for many quantities, the Census has given us data for County Divisions, Cities, Towns, and Census Traits. We have the raw material for looking inside counties, inside large cities, etc. and we can use it to move far toward a

fit PLUS residuals

position. Once there, we can, today:

---

*to NE.

- map the raw residuals
- map the residuals shrunken for sampling variation
- map the smoothed residuals

and tomorrow we should be able to combine shrinking for sampling variation with smoothing.

Even if the Census had not been as kind to us as it has, we could move a major step toward

fit PLUS residuals

using only county data. 3000-odd counties are a lot. We have no need at all to stick to patch maps, we can do much better.

The story is always the same:

- somewhat more calculation
- more difficult explanation
- a little greater uncertainty
- all about much more useful numbers and pictures.

As so often, the gain is likely to be great.

HAVE I BEEN EXTREME?

Some of you may find what I have been saying very extreme. Too bad, for I have really been quite conservative. Many of you know my Harvard friend and colleague Frederick Mosteller, most of you know who he is. He is conservative enough to have been chosen, in a very Republican Administration, Vice Chairman of the President's Commission on Federal Statistics. But he asserts, and I am glad to join in his assertion, that we have a real responsibility, AS STATISTICIANS, to guess when only guessing is possible. His practical suggestions on how to learn to do such things will appear in a forthcoming book of readings and writings. I urge you to read them.

WHY?

Surely there must be reasons other than those I have suggested why we have not moved further ahead into the uncomfortable, if that is where the most valuable things are to be done. Surely there must, and I wish I could be surer what they are. Some, as we noted, are matters of organizational pressure, actual or perceived.

One largish one I can identify, I think. It relates to the use of

fit PLUS residual

as a way of life, where the fit may come wholly from the data before us wholly from other experience (data or crystal ball) or partly from one and partly from the other.

I look upon the process of making such a separation AND then looking carefully at BOTH parts, separately (and, if it helps, together) as a standard process, empirical and exploratory in nature; as not necessarily connected with any notion of best estimation -- or even any notion of estimation, as a convenience in making contact with our data, and not as something that necessarily teaches us about laws of nature or even about continuing regularities. Accordingly I am willing to try one fit in the Northeast -- and another in the West; or one fit in more metropolitan counties and another in the open country. If one is willing to follow where the

data leadeth, and sit loose to the meaning and constancy of one's fits, the

fit PLUS residual

approach can do much for any of us, just as the reverse attitude can get any of us in trouble.

As statisticians, we have a responsibility to recognize patterns, to make behavior describable in words, to reduce that which is left over, as far as is reasonable, to undescribed irregularity. The price of making more of what happens describable is making what is described somewhat more complicated. Not all our clients and users will be realistic enough to like this, but enough can be to nucleate understand and spread the more realistic view. (Last Thursday, the Wall Street Journal also reported [4] the Federal Trade Commission's sadness that consumers might not understand a measure of air conditioner efficiency, and their rather plaintive inquiry as to whether it might not be possible to put it in watts. I believe enough consumers will be quite clear about efficiency measures to make their use generally effective, and I believe that our consumers can and will learn to live with a slightly more complicated world.)

Clearly what I have said today has drifted somewhat from what I thought I might say. Unsurprisingly enough, it has drifted toward things I understand in more detail. But the key flavor is the same:

> *If what is really needed is harder to measure or harder to explain, we still need to measure and explain it.*

And my deep conviction that we are failing to measure important things, either because of pride in our measurement techniques or because of fear of imprecision being interpreted as error, is undisturbed and evergrowing. To refuse to try to measure something because there is no good frame can easily be "hubris", the kind of pride that leads to a fall. If too many of us are too proud of our measurement techniques, or show the wrong degree of bravery in reporting, we may all fall together.

REFERENCES

[1] *The National Halothane Study* 1969 (ed. J. P. Bunker, W. H. Forrest, F. Mosteller, L. D. Vandam) National Institutes of Health, ix + 415 pp.

[2] (tables) T. J. Mason, F. W. MacKay, 1974(?) U. S. *Cancer Mortality by County:* 1950-1969. (DHEW Publication No. (NIH) 74-615), National Cancer Institute, U. S. Department of Health, Education, and Welfare, vi + 729 pp.
(patch maps) T. J. Mason, F. W. MacKay, R. Hoover, W. J. Blot, J. F. Fraumeni, Jr., 1975(?) *Atlas of Cancer Mortality for U. S. Counties:* 1950-1960 (DHEW Publication No. (NIH) 75-780, National Cancer Institute, U. S. Department of Health, Education, and Welfare, xi + 103 pp.

[3] New Economics: Surge of Price Rises Stirs Debate on the Causes, *Wall Street Journal,* p. 1 and 19, (see end of story on page 19) 22 August 1975.

[4] FTC Seeks to have Energy Date included in Air Conditioner advertising. *Wall Street Journal,* 21 August 1975, p. 2.